Outils d'accès à des ressources linguistiques

Laurent Romary

Chapitre 8

Outils d'accès à des ressources linguistiques

8.1. Introduction

Il est difficile de concevoir l'évolution de l'ingénierie linguistique sans l'existence concomitante de ressources linguistiques qui en alimentent les processus. En fonction de la nature de ces ressources (lexiques, dictionnaires, transcription de dialogues, textes éventuellement annotés), elles sont susceptibles de participer à différents stades de la conception d'un système de traitement des langues naturelles :

- Paramétrisation d'un processus de traitement tel que l'analyse syntaxique (reposant par exemple sur un lexique associé à un ensemble d'arbres élémentaires dans le cas d'un analyseur LTAG) ;
- Apprentissage de modèles stochastiques à partir de textes annotés (par exemple pour l'étiquetage morphosyntaxique) ;
- Validation d'un processus particulier après sa conception. Dans ce cas le corpus annoté va servir de référence par rapport à laquelle les résultats du processus considéré seront évalués [ADDA 1999] ;

Plus largement, il ne faut pas limiter le champ de notre analyse à la seule problématique de l'ingénierie linguistique si l'on considère que les mêmes ressources sont de plus en plus utilisées en amont pour accompagner le travail de modélisation linguistique, à la fois à des fins d'observation (recherche d'occurrences de mots, de structures ou de phénomènes sémantique ou pragmatique particuliers) et à des fins, là aussi, de validation de modèles. De fait, c'est ce domaine qui a le premier posé le problème de l'accès à de grandes quantités de ressources linguistiques et fait émerger les premiers concepts généraux à ce sujet.

Cependant, on ne peut que constater la très faible diffusion à la fois des ressources et des outils d'une équipe de recherche à une autre, ce qui semble être un frein important au développement plus rapide, et nécessaire, du domaine du traitement automatique des langues. Ce chapitre essayera donc à la fois de fournir un panorama des réponses existantes dans le domaine de l'accès aux ressources linguistiques, mais aussi des éléments d'analyse sur les problèmes rencontrés à l'heure actuel.

8.2. Un premier diagnostic

8.2.1. Facteurs limitant l'accès aux ressources linguistiques

Il y a plusieurs façons d'analyser l'absence d'une large diffusion de ressources linguistiques au sein de la communauté du traitement automatique des langues. Un premier facteur en est probablement le caractère extrêmement ciblé des recherches menées dans les différentes équipes et donc de la spécificité des ressources qui sont parallèlement produites. Il faut cependant relativiser. À une extrémité, on peut admettre qu'une base terminologique sur les bactéries du lait sera difficilement utile à une large communauté de chercheurs. Éventuellement, La transcription d'un corpus de dialogue oraux peut elle aussi, en fonction du contexte du recueil, présenter des idiosyncrasies difficilement transposables. Un tel corpus peut être pourtant utile à l'étude plus générale de phénomènes discursifs par exemple, et les difficultés que l'on rencontrera seront plus liées aux spécificités des règles de transcription adoptées. Enfin, il est clair que pour des textes écrits, ce facteur joue beaucoup moins, et il est surtout important d'être à même d'une part d'identifier correctement la source et le contenu du texte et d'autre part d'adopter des formats normalisés de représentation de la structure (divisions et paragraphes notamment) et des quelques éléments de surface que la majorité des compilateurs repèrent. Les seules réelles difficultés que l'on peut rencontrer correspondent aux annotations que chaque équipe souhaite ajouter à ces textes, qu'il s'agisse d'étiquetage morpho-syntaxique, de découpage en phrase ou de repérage de phénomènes de plus haut niveau comme la référence par exemple [Bruneseaux 1997]. Une bonne partie des réponses à ce type de difficultés réside dans l'adoption d'un cadre de normalisation des représentations des différents niveaux de transcription ou d'annotation de ces ressources, comme on a pu le voir au chapitre précédent.

Une autre catégorie de facteurs est liée au temps nécessaire à la création de ressources linguistiques et donc au coût intrinsèque auquel celles-ci correspondent. La compilation d'une base de texte de qualité, comme la mise en œuvre d'un lexique de taille raisonnable et contenant des informations validées d'un point de vue linguistique correspondant souvent, pour les équipes concernées, à un travail de longue haleine qui a accompagné le développement de l'expertise scientifique associée. Il est alors difficile pour ces équipes de donner un accès intégral à leurs données car elles ont l'impression soit de se faire piller, soit tout simplement de perdre une partie de ce qui fait leur reconnaissance scientifique. Ce dernier argument s'avère dans les faits trompeur puisque

les équipes qui se sont engagées dans la voie de la diffusion large de leurs ressources et des méthodologies associées ont bénéficié d'un regain de renommée non négligeable (on peut citer par exemple la diffusion du corpus *Map Task* par le HCRC à Edimbourg [Anderson 1991]), puisqu'il s'agit bien de la compétence qui est alors évaluée et non pas le fait de posséder une ressource.

On peut citer enfin, le problème des droits d'auteurs ou d'éditeurs associés à l'usage et la diffusion des ressources les plus récentes. Le développement de l'Internet a conduit la plupart des éditeurs à adopter une attitude de replis qui rend la gestion de ces problèmes particulièrement complexes, notamment à l'échelle d'une équipe particulière. Le développement de structures commerciales comme l'ELRA (European Linguistic Resource Association), même si elles peuvent avoir leur place dans le développement industriel de l'ingénierie linguistique, ne répond pas sur ce point aux besoins des laboratoires académiques. Il est indispensable que les organismes de recherche (Universités, CNRS, BNF etc.) prennent eux-mêmes les choses en main pour en particulier faire pression sur le législateur et favoriser l'émergence d'un cadre spécifique où ces ressources pourront être utilisées plus librement.

8.2.2. Facteurs limitant le développement et la diffusion d'outils réutilisables

L'absence d'un cadre unifié permettant de diffuser largement les ressources linguistiques disponibles dans les laboratoires explique en grande partie la faible diffusion d'outils au sein de la communauté du traitement automatique des langues. Ainsi, si l'on observe un certain nombre de plate-formes d'analyse syntaxique, on observe que même aux niveaux les plus élémentaires, les lexiques sur lesquels elles reposent ne sont pas du tout inter opérables. La conséquence immédiate est que tout chercheur qui s'implique dans l'utilisation d'un outil particulier est contraint sur le long terme de se tenir au format qui lui est imposé et que se forment ainsi une communauté qui a d'une part du mal à se remettre en cause et d'autre part extrêmement fragmentée. On peut malgré tout reconnaître les apports d'initiatives récentes telles que Multext [BEL 1995], qui ont pu définir un format unifié de description des étiquettes morpho-syntaxique, à même de conduire progressivement à l'implantation d'outils génériques en la matière.

Au final, le bilan que l'on peut tirer de cette analyse est presque plus sociologique que scientifique. Considérant que l'adoption d'une démarche de normalisation des données et du développement des logiciels en traitement automatique des langues doit correspondre à un investissement non négligeable pour les laboratoires concernés, mais qu'a contrario le gain pour l'ensemble de la communauté résultant de cette mise en commun des compétences peut être de faire de réelles avancées de recherche, auront nous cette capacité à rendre prioritaire l'intérêt du groupe devant les intérêts particuliers. L'objectif de ce chapitre est dans ce cadre de faire le point sur les acquis dans le domaine des outils d'accès aux ressources linguistiques tant d'un point de vue méthodologique que

technologique et de présenter une intégration effective de ces acquis au travers de réalisations concrète auxquelles les auteurs de ce chapitre ont participé.

8.3. Méthodes d'accès aux ressources

8.3.1. Accès au contenu linguistique

Concordances et statistiques lexicales élémentaires

L'accès en ligne à des ressources linguistiques a longtemps correspondu au besoin des lexicographes d'observer des occurrences de mots en contexte et s'est donc souvent ramené à la notion de *concordance*. Il s'agit ainsi de présenter une suite de lignes de contextes, centrées sur la forme recherchée et éventuellement triées en fonction des formes apparaissant à gauche ou à droite de celle-ci. De façon à affiner l'observation lorsque le nombre de contexte est plus important, on adjoint quelques outils statistiques élémentaires permettant notamment de comptabiliser les fréquences des différentes formes rencontrées dans le corpus utilisé.

Recherche de collocations dans un texte

Lorsque l'on cherche à décrire le fonctionnement sémantique de mots sur la base du comportement effectif de ceux-ci dans des textes, la notion de *collocation* est particulièrement importante. L'idée est de mettre en évidence les groupements lexicaux dont la co-occurrence est significative d'une régularité de sens à l'intérieur des mots composant ce groupement. Pour cela on utilise le plus souvent des tests statistiques qui reposent à peu près tous sur les mêmes concepts. On part ainsi *d'un mot pôle*, dont on va étudier les occurrences à l'intérieur d'un *corpus de référence*. On extrait ainsi un ensemble de concordances correspondant à des *fenêtres* d'observation de taille fixe ou variable autour du mot pôle, pour lesquelles on recueille l'ensemble des formes ou parfois des lemmes. Pour chacune de ces formes, on évalue leur fréquence d'occurrence relativement au corpus de référence pour déterminer si leur présence est significative ou non.

Plus concrètement, il existe tout un ensemble de tests issus soit de la théorie de l'information, soit du domaine des statistiques. Ainsi, pour mettre en évidence la proximité sémantique de deux termes, l'information mutuelle I(a,b) évalue la force du rapport entre les fréquences relatives d'apparition de deux mots a et b dans un même contexte au regard de leurs fréquences propres d'apparition dans le corpus [CHURCH 1991]. Ce score symétrique est souvent opposé à la formule de l'écart réduit ER(a) [MULLER 1977] (ou *Z-score* dans les milieux anglo-saxons [BERRY-ROGGHE 1973]) qui estime la pertinence de l'écart entre la fréquence absolue observée d'un mot a dans un sous-corpus (formé généralement de l'ensemble des contexte autour du mot pôle b) et la valeur estimée de cette fréquence étant donné le corpus de référence. L'avantage de ce dernier test est notamment que l'on peut adapter sa formulation au contexte d'étude en

changeant la loi de probabilité de référence (binomiale, hypergéométrique ou Poisson). Dans les deux cas, il peut être nécessaire de rejeter les résultats relatifs aux mots dont le nombre d'occurrence dans le corpus de référence est trop faible. Pour mémoire, on résume ci-dessous les formules de ces deux scores, en adoptant une notation unifiée :

Soit N la taille du corpus de référence (nombre de mots), X la fréquence absolue (nombre d'occurrences) du mot à observer a, Y la fréquence absolue du mot pôle b, x le nombre d'occurrence de a dans les fenêtres d'observation, f la taille des fenêtres d'observation (en mots) et n la taille totale du sous-corpus d'observation correspondant aux fenêtres (n=Y*f). L'information mutuelle pour a et b s'exprime ainsi :

$$I(a,b) = \log_2 \frac{P(a,b)}{P(a)P(b)}, \text{ avec } P(a) = \frac{X}{N}, \; P(b) = \frac{Y}{N} \text{ et } P(a,b) = \frac{x}{N}$$

L'écart réduit, dans le cas d'une loi binomiale, s'exprime quant à lui de la façon suivante :

$$ER(a) = \frac{f\_observée \; - f\_estimée}{ecart\_type} = \frac{x - \frac{nX}{N}}{\sqrt{n \frac{X}{N} \frac{(N-X)}{N}}}$$

Il existe par ailleurs un certain nombre de tests permettant de comparer plusieurs hypothèses lexicographiques. On peut en particulier citer le t-score [CHURCH 1991], qui permet de mettre en évidence la différence entre deux structures linguistiques et le test du Khi deux, qui identifie l'indépendance ou non de plusieurs hypothèses.

Dans les tests présentés ci-dessus, la taille des fenêtres joue un rôle significatif sur la pertinence et la nature des résultats obtenus. Comme on peut en avoir l'intuition, des fenêtres de trop petite taille vont tendre à renforcer les phénomènes de co-occurrence syntaxique (par exemple la liaison entre l'article et le nom), notamment dans le cas des expressions figées. Inversement, des fenêtres plus larges engendrent un effet de dilution sur les associations sémantiques observées. Une taille de 20 formes semble fournir de bons résultats [VALCESCHINI-DEZA 1999]. Par ailleurs, pour préserver la cohérence des résultats avec la structure logique des textes analysés, il peut être pertinent de contraindre les fenêtres d'analyse à respecter les frontières de phrase, de paragraphe ou de chapitre. Cependant, ceci complique les calculs puisque les différentes fenêtres d'observation n'ont plus une taille fixe et d'autre part, il est nécessaire (cf. infra) de disposer de cette information structurelle à l'aide d'un codage particulier.

De la même façon, il faut prêter une attention particulière à la nature des objets observés. Classiquement, on travaille directement sur les graphies, correspondant aux formes fléchies, mais on peut aussi s'intéresser aux lemmes, moins sensibles aux

variations morpho-syntaxiques [RASTIER 1994]. Cependant, il est important de toujours conduire une étude préliminaire sur les graphies de sorte à s'assurer qu'on ne laisse pas de coté des phénomènes sémantiques spécifiques qui seraient liés au genre ou au nombre par exemple. Il faut aussi noter l'importance des signes de ponctuation dans l'organisation sémantique d'un texte. Les tests ci-dessus donnent des résultats intéressant quand il sont appliqués à l'ensemble des graphies d'un document [BOURION 1998].

Enfin, il faut noter l'importance essentielle du corpus de référence au sein duquel les observations lexicales sont effectuées. Comme on l'a vu, la pertinence des résultats n'est que relative et doit être envisagée au regard du genre (journalistique, littéraire, technique, etc.), du domaine et de la période temporelle couverts par le corpus que l'on aura sélectionné. Plus le corpus sera homogène, plus les résultats obtenus seront facilement interprétables. On pourra ainsi observer des idiosyncrasies d'auteurs sur l'usage de tel ou tel lexique si on se limite aux œuvres de celui-ci [RASTIER 1995]. À l'inverse, un corpus extrêmement hétérogène sera trop bruité pour que les tests statistiques puissent véritablement mettre en évidence des phénomènes significatifs. Même si on se fixe comme objectif d'observer des régularités d'ordre général pour une langue donnée, il est indispensable de fixer des paramètres stables minimaux comme le genre.

En guise de bilan, on constate que les outils d'accès aux contenus des textes sont seulement en cours de définition et il n'est guère possible à l'heure actuelle d'envisager la mise en œuvre de processus complètement automatiques pour construire par exemple un véritable lexique sémantique à partir d'un corpus de textes. Cependant, le fait que l'on dispose à la fois de bases de textes de plus en plus importantes et de moyens de calculs de plus en plus performants font qu'il s'agit là d'un domaine de recherche particulièrement actif. De nouveaux champs d'investigations émergent ainsi, que ce soit pour l'étude de phénomènes textuels plus larges tels que la thématique (en prenant par exemple des sous-corpus d'étude formés de large portions de textes ou même de textes complets), ou pour sortir du strict cadre monolingue et étudier la possibilités de faire émerger (semi-) automatiquement des équivalents de traduction à partir de corpus de textes parallèles ou comparables [DAGAN 1994].

8.3.2. Accès à la structure des documents

Bilan rapide des apports d'XML

Comme on l'a vu au chapitre précédent, l'arrivée de XML a joué un rôle important dans l'évolution des concepts sous-jacents à la notion de ressource linguistique. En particulier, XML peut être considéré, par sa filiation à SGML et donc à toute une branche d'activités liées à l'industrie de l'édition scientifique et technique, comme un langage de représentation documentaire, mais aussi, conséquence de son positionnement récent dans le domaine de l'échange d'information sur l'Internet, comme un véritable langage de description de données. C'est ainsi que l'on assiste à un rapprochement

possible entre la représentation de documents textuels, qui possèdent une structure relativement plate reposant sur un balisage réduit et faiblement contraint, et l'organisation de ressources telles que les dictionnaires qui reposent sur une structure hiérarchique fortement contrainte par un modèle de données relativement fin (voir l'exemple présenté figure 8.1). Classiquement, on considérait que cette dernière classe relevait du domaine des bases de données alors que la première ne présentait de l'intérêt que par son contenu brut, certains chercheurs regardant même le balisage comme secondaire, voire inutile [SINCLAIR 1991].

D'un point de vue plus opérationnel, la disparition du concept strict de document valide qui, dans la norme SGML, imposait que toute instance de document soit conforme à une DTD[1] qui lui était systématiquement associée, pour une vision plus souple autorisant un document à n'être que bien formé, c'est-à-dire pour lequel l'organisation intrinsèque des balises dans l'instance peut être reconstruite de façon univoque, permet de concevoir de nouvelles architectures de gestion et de transmission de l'information. Il est en effet possible, à partir d'un document primaire d'origine (validé au regard d'une DTD connue), de n'en transmettre qu'une partie qui soit pertinente pour un traitement donné ou suite à une requête d'un utilisateur. Ainsi, un paragraphe de roman, une réplique de théâtre ou l'entrée d'un dictionnaire peuvent être considérés comme des documents XML parfaitement valides qu'un processus logiciel pourra manipuler ou transmettre. À l'inverse, des documents ou parties bien formées de documents issus de sources différentes peuvent être re-combinées pour former un nouveau document. Par exemple, on peut construire une entrée lexicale réunissant des descriptions extraites de dictionnaires existants, de différentes concordances issues d'une base de textes et d'informations plus formelles à des fins de traitement automatique. Cette interopérabilité des données rend même la description de celles-ci beaucoup plus souple, puisque l'on peut envisager de définir des modèles de documents plus restreints (dédiés par exemple aux structures génériques d'un texte, aux entrées de dictionnaire, aux références bibliographiques, etc.) que l'on combine ensuite en les distinguant à l'aide du mécanisme des espaces de noms[2].

Enfin et surtout, la force de XML est de devoir son succès à l'Internet et de disposer ainsi d'une plate-forme extrêmement vaste d'utilisateurs travaillant sur les mêmes bases technologiques. Par effet d'entraînement, la communauté de l'ingénierie linguistique,

---

[1] La DTD (Document Type Definition ou définition du type de document) décrit la structure abstraite d'un document sous la forme des enchâssements valides d'éléments ainsi que les attributs que ceux-ci peuvent porter.

[2] Les espaces de noms [Namespaces 1999] permettent d'identifier l'appartenance d'une balise donné à un groupe de référence et ainsi d'éviter de possible ambigüités lors de la combinaison de documents d'origines différentes. À titre d'exemple, l'entrée de dictionnaire de la figure 8.1 pourrait être préfixée en `<dico:entry><dico :form>…</dico:entry>` pour peu que l'espace de nom 'dico' ait été décrit au début du document correspondant.

même pour ses éléments les plus réticents, se doit d'adopter ce format de représentation pour bénéficier des nombreux développements logiciels qui accompagnent le déploiement de ce standard.

Modèle hiérarchique et chemins d'accès

D'un point de vue formel, XML peut être associé à un modèle de donné arborescent où chaque élément est représenté par un nœud dont les fils sont formés des éléments compris à l'intérieur de celui-ci et éventuellement des blocs textuels qu'il contient. On a ainsi représenté figure 8.1 la structure arborescente associée à l'entrée de dictionnaire qui y est décrite.

```
<entry>
 <form>
  <orth>table</orth>
 </form>
 <gramGrp>
  <pos>n</pos>
  <gen>f</gen>
 <gramGrp>
 <def>Pièce de mobilier…</def>
 <eg>Une table de cuisine</eg>
</entry>
```

`entry` : entrée d'un dictionnaire
`form` : indications morphologiques
`orth` : forme orthographique
`gramGrp` : indications grammaticales
`pos` : partie du discours
`gen` : genre
`def` : définition
`eg` : exemple

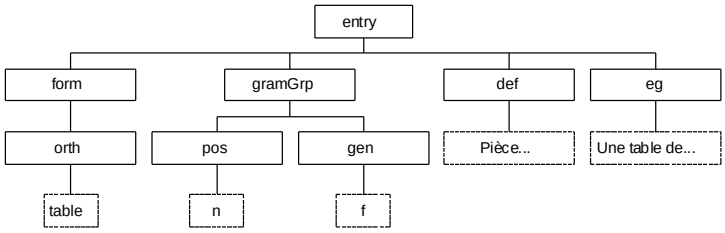


Figure 8.1 : *Description d'une entrée d'un dictionnaire en XML (directives de la TEI) et représentation arborescente.*

Dès lors, on dispose d'un mécanisme de désignation de tout nœud d'une telle hiérarchie en décrivant le chemin qui permet d'y accéder. Le langage correspondant repose sur la définition d'un certain nombre d'axes de parcours[3] le long d'un document

[3] `child` (les fils d'un nœud), `descendant` (la fermeture transitive de `child`), `parent` (le père d'un nœuds), `ancestor` (la fermeture transitive de `parent`), `following-sibling` (les frères à droite), `preceding-sibling` (les frères à gauches), `following` (les nœuds suivant le nœud

qui, associés à l'expression de contraintes sur le contenu, les attributs ou la position des nœuds obtenus, permet d'exprimer des requêtes plus ou moins complexes. On peut ainsi retrouver l'entrée de dictionnaire utilisé dans notre exemple à l'aide du chemin suivant :

```
/descendant::entry[child::form/child::orth='table']
```
[4]

où l'on a d'une part l'indication que l'on recherche un nœud de type `entry` comme descendant quelconque due la racine du document courant et d'autre part la contrainte qu'il contienne un nœud `form` qui contient lui-même un nœud `orth` dont le texte est égal à la chaîne de caractères requise.

Outils d'accès à la structure d'un document

La mise en œuvre de XML d'un point de vue logiciel repose sur la définition d'interfaces de programmation normalisées (ou API, *Application Programming Interface*) permettant à tout concepteur de système de s'appuyer de façon transparente sur des librairies conformes à ces interfaces. Il existe ainsi une interface de base permettant de gérer un document XML vu comme une arborescence de nœuds conformément au modèle exposé au paragraphe précédent [DOM 1998]. Cette interface permet notamment de parcourir l'arborescence à partir d'un nœud donné dans différentes directions (père, liste des fils, premier fils, dernier fils, etc.) et d'accéder à ou de modifier les propriétés d'un document ou d'un nœud (par exemple ses attributs).

De façon duale, il existe une interface de programmation simplifiée pour XML (SAX, A simple API for XML, [SAX 1998]) qui, au lieu de s'appuyer sur la description complète de l'arbre associé à un document XML, va reposer sur un mécanisme d'évènements associés à chaque type d'objet (début de document, balise ouvrante ou fermante, contenu textuel etc.) rencontré lors du parcours du document. Cette interface, bien que plus limitée, permet d'une part d'accéder plus rapidement à une information particulière dans un document et d'autre part de s'affranchir des contraintes de place en mémoire liées à la manipulation de très gros documents.

Alors que les éléments logiciels présentés ci-dessus sont destinés aux informaticiens qui vont les intégrer dans des applications plus larges, XML s'accompagne, à un deuxième niveau, de recommandations apparentées qui permettent de manipuler des structures par filtrage et recombinaison des nœuds que celles-ci contiennent. Le langage XSLT [XSLT 1999], qui repose sur les chemins XPath et qui était destiné initialement à décrire des feuilles de style associées à des documents XML est un véritable langage de manipulation de tels documents. Décrit lui-même en XML, il permet à des non spécialistes d'exprimer des règles simples de restructuration de leurs documents (par

---

courant), `preceding` (les nœud qui précèdent le nœud courant), `self` (le nœud courant), `descendant-or-self` et `ancestor-or-self`.

[4] Il existe une forme abrégée qui permet d'exprimer le même chemin ainsi : `//entry[form/orth='table']`

exemple pour passer d'une DTD à une autre) ou à niveau plus basique d'effectuer des présentation simples de ceux-ci. Pour illustrer ceci, la figure 8.2 contient des règles simples de transformation de l'exemple de la figure 8.1 en HMTL accompagné du résultat final que l'on visualiserait sur un navigateur Internet standard.

`<xsl:template match="entry">`
`  <p><xsl:apply-templates/></p>`
`</xsl:template>`
`<xsl:template match="gramGrp">`
`  (<xsl:value-of select="."/>)`
`</xsl:template>`
`<xsl:template match="eg">`
`  <i><xsl:value-of`
`  select="."/></i>`
`</xsl:template>`
	Chaque entrée est mise dans un paragraphe autonome ;
	
	Les indications grammaticales sont mises entre parenthèses
	
	Les exemples sont indiqués en italiques.

table(nf)Pièce de mobilier…*Une table de cuisine.*	

Figure 8.2 : *une feuille de style XSL et son application à l'exemple de la figure 8.1.*

Bilan

La venue de XML a permis de répondre en grande partie au problème d'identifier une couche de base permettant de représenter la structure des ressources linguistiques utilisées en traitement automatique des langues. Bien plus, comme on le verra dans la section suivante, ce métalangage peut servir à définir les protocoles de référence pour tous les échanges de données entre modules de traitement dans des architectures plus complexes. Il reste à définir des interfaces de programmation normalisées dédiées plus spécifiquement à l'ingénierie linguistique qui permettraient par exemple de manipuler des lexiques de façon transparente.

8.3.3. Mise en œuvre d'un réseau de serveurs ressources linguistiques

L'étape ultime dans la définition d'outils d'accès à des ressources linguistiques et de rendre ceux-ci les plus transparents possibles de sorte qu'un utilisateur final n'ait qu'à formuler une requête intuitive pour obtenir les résultats qu'il recherche. Nous allons ainsi présenter dans cette section les grandes lignes de la définition de serveurs de ressources linguistiques en considérant plus particulièrement le cas de leur mise en réseau. En effet, alors qu'il existe un certain nombre d'outils d'accès à des bases de textes fonctionnant soit sur des mécanismes relativement simples (par exemple Concordance[5]) ou intégrant

[5] http://www.rjcw.freeserve.co.uk/

des fonctionnalités, notamment statistiques, particulièrement élaborées (Thief d'Etienne Brunet), il s'agit en général de logiciels fermés. Ils reposent sur leur propre logique et sont de ce fait d'une part extrêmement lié à des formats de données propriétaires et d'autre part difficilement utilisable en dehors du cadre étroit d'utilisation qui a conduit à leur conception.

Organisation générale du réseau

La figure 8.3 présente l'architecture globale que peut prendre un réseau de serveurs de ressources linguistique. Pour en analyser plus finement le fonctionnement, on peut considérer l'activité d'un d'utilisateur qui s'y connecterait et considérer que celle-ci peut se scinder en un certain nombre de phases distinctes :

- Connexion au serveur et identification, au cours de laquelle un utilisateur référencé charge son profil de session ;
- Sélection des serveurs vers lesquels des requêtes vont être transmises. L'utilisateur peut ainsi souhaiter ne travailler que sur un nombre réduit de serveurs spécialisés ou au contraire interroger largement l'ensemble des ressources disponibles ;
- Envoi d'une requête d'identification de ressources et sélection de celles-ci, qui vont permettre à l'utilisateur de se construire itérativement un corpus de travail ;
- Envoi d'une requête sur les contenus des ressources sélectionnées, qui correspond en quelque sorte à la phase effective de travail.

Ce scénario nous conduit à identifier deux types de classes de serveurs qui peuvent éventuellement se réaliser en un même serveur effectif : d'une part des serveurs d'accès qui sont les points d'entrée sur le réseau pour les utilisateurs, et d'autre part des serveurs de ressources, qui interprètent effectivement les requêtes des utilisateurs.

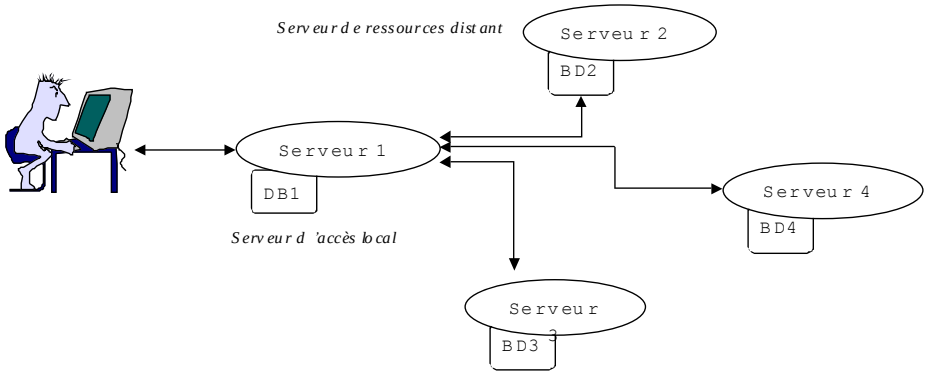


Figure 8.3 : *structure générale d'un réseau de serveurs de ressources linguistiques.*

Serveur d'accès

Le serveur d'accès joue plusieurs rôles essentiels pour le bon fonctionnement du scénario présenté ci-dessus. Tout d'abord, il lui incombe de gérer une base d'utilisateurs, qui sont éventuellement associés à un niveau de droit d'accès reconnus globalement par le réseau. Il doit tenir à jour une liste des serveurs accessibles sur le réseau en fonction des informations fournies par le gestionnaire de réseau (cf. infra). Enfin et surtout, il a pour fonction de rediriger vers ces différents serveurs toute requête de la part de l'utilisateur et combiner en retour les ensembles de résultats fournis par chacun d'eux.

Serveur de ressources

Le serveur de ressources est l'unité de traitement qui procure un accès souple et rapide aux données linguistiques en interprétant les requêtes de l'utilisateur transmises par le biais du réseau. L'une des fonctions majeures à implanter au niveau d'un tel module est la mise sous forme d'une réelle base de données des ressources disponibles. Il s'agit en général de réaliser une structure inversée qui indexe les différentes unités des textes (mots et ponctuations) et gère, parallèlement à cet index, un ensemble d'attributs associés à chaque unité (par exemple les étiquettes morphosyntaxiques). Bien que peu de moteurs existants aient réellement intégré une solution élégante sur ce point (cf. infra à propos du langage de requête), la plupart proposent une indexation des structures SGML quand celles-ci existent. Deux options sont alors envisageables : soit, et ce n'est probablement pas à retenir à terme, considérer les balises comme des unités du texte, soit intégrer la structure arborescente induite par les balises dans la définition du chemin d'index, ce qui permet des comparaisons rapides de position des éléments aux différents niveaux structurels.

Langage de requête

Comme élément de liaison entre l'utilisateur et le serveur de ressources, le langage de requête doit théoriquement combiner deux contraintes opposées, c'est-à-dire, une certaine puissance d'expression correspondant aux capacités du serveur, et par ailleurs une relative simplicité pour que l'utilisateur soit capable de se l'approprier. De fait, nous considérons que ce dernier aspect n'est pas prioritaire si on considère que l'on doit offrir à l'utilisateur des interfaces lui permettant de formuler des requêtes sans qu'il connaisse nécessairement les spécificités syntaxiques du langage sous-jacent.

On peut ici recenser les principales possibilités offertes par un langage de requête sur des ressources textuelles en les illustrant d'exemples tirés du système développé à l'Université de Stuttgart (IMS Corpus Workbench [CHRIST 1994]) :

- Interrogation sur des suites de mots ("pris" "en" "grippe";) ;
- Utilisation d'un opérateur de portée permettant de contraindre l'intervalle d'occurance d'une forme donnée ("pris" []{2,8} "en" "grippe"; *le verbe est séparé de son complément* en grippe *par un nombre de mots compris entre 2 et 8*) ;

- Possibilité d'exprimer des formes incomplètes à l'aide d'opérateurs logiques ou d'expressions régulières ("pris | prise | prises" "en" "grippe";) ;
- Possibilité d'adjoindre d'autres contraintes sur les formes en fonctions d'attributs complémentaires, par exemple les parties du discours ([word="pris | prise | prises" et pos="V"] "en" "grippe") ;
- Contraindre l'étendu d'une requête à un contexte structurel particulier, tel qu'exprimé par un balisage XML par exemple ("pris" "en" "grippe" within s; *identification de la séquence à l'intérieure d'une phrase balisée par <s>*) ;

Utilisation de XML comme protocole général de communication

XML, comme langage général de représentation de données semi-structurées, peut prendre en compte de nombreux types différents d'information. Il est alors tentant, en particulier pour uniformiser les traitements de bas niveau sur les données échangées au sein du réseau de serveurs ou en direction du client, d'utiliser XML de façon exclusive à cet effet. On peut ainsi définir un ensemble de DTD qui vont décrire [ELAN 1999] les données utilisateurs (paramètre d'identification, langue préférentielle etc.), l'espace de travail dynamique au cours d'une session (intégrant les sélections de ressources et les résultats temporaires), et surtout le protocole de transmission des requêtes et de récupération des résultats. En dehors d'une simplification de mise en œuvre, l'avantage d'utiliser XML à tous les niveaux est bien évidemment de garantir une compatibilité des données pour par exemple intégrer les résultats d'une requête dans l'espace de travail).

Mécanisme de cache

Dans un système client-serveur tel que nous le décrivons ici, il n'est pas question de télécharger au niveau du client l'ensemble des résultats associés à une requête donnée. À l'inverse, il faut éviter de laisser toute la charge du traitement au serveur de ressource distant et de relancer une requête dès que l'on passe d'un résultat élémentaire à un autre. Il est donc nécessaire de mettre en œuvre un système de cache, c'est-à-dire de mémorisation temporaire des résultats disponibles de sorte à gérer harmonieusement les flux de données entre le serveur distant et le client. Dans une architecture en réseau, il est ainsi possible de considérer trois niveaux de cache :

- Au niveau du serveur distant pour la gestion de l'ensemble des résultats correspondant à une requête ;
- Au niveau du serveur d'accès pour cumuler les résultats issus des différents serveurs distants ;
- Au niveau du client lui-même pour garder en mémoire les résultats les plus récents présentés à l'utilisateur et éviter ainsi des connexions inutiles.

Ce mécanisme de cache ne fonctionne que si, au niveau du protocole de requête et de renvoi des résultats, on a intégré une gestion des flux de données. Celle-ci doit fournir, au niveau des requêtes, des marqueurs de début, de fin ou de quantité maximale désirée

("je souhaite au plus 50 réponses à partir de la 350^{e}") et, au niveau des résultats, le positionnement effectif de ces marqueurs dans l'espace total des résultats possibles.

Gestionnaire de réseau

Un réseau de serveurs de ressources linguistiques, comme tout réseau spécialisé de serveurs d'informations, ne peut fonctionner sans un minimum de centralisation. Dans le cadre du réseau ELAN, un gestionnaire central de réseau a ainsi été mis en œuvre. Ce gestionnaire ou NMU (*Network Management Unit*/Unité de Gestion du Réseau) tient à jour une liste des serveurs rattachés au réseau, comprenant le nom du serveur, sont adresse physique ainsi qu'un court descriptif de sa spécialité (langue, genre ou période privilégiée). L'ajout ou le retrait du réseau ne se fait que via la NMU qui informe à son tour chacun des serveurs des autres serveurs reconnus par le réseau. Là encore, un mécanisme de cache local à chaque serveur permet à l'ensemble du réseau de fonctionner, même si la NMU est temporairement hors service.

8.4. Définition d'une plate-forme éditoriale pour une base terminologique multilingue

8.4.1. Contexte général

Il peut être intéressant de clore ce chapitre en présentant une perspective différente sur l'accès aux ressources linguistiques. L'objectif est ici de considérer non pas l'accès en lecture à des ressources, mais le problème de leur édition (ou annotation) en ligne. Au-delà du domaine spécifique que nous allons présenter, lié à un projet très spécialisé, l'édition en ligne de ressources linguistiques est pour nous l'un des grands axes de développement des travaux dans le domaine de l'ingénierie linguistique pour les années à venir. Il s'agit en effet d'envisager des méthodes et des techniques qui permettront d'ajouter à des ressources données des annotations supplémentaires partagées entre spécialistes ou à destination d'autres populations intéressées par ces informations. Du point de vue des applications, on peut tout aussi bien penser à des traductions ou commentaires sur des documents originaux (transcriptions en ethnolinguistique, manuscrits anciens, littérature etc.), qu'à des séances de travail en commun en classe de littérature. Suivant le type d'environnement, il sera nécessaire de se poser des questions d'accessibilité de tout ou partie des ressources et de leurs annotations, de certification des nouvelles informations, ou de synchronisation de celles-ci quand plusieurs personnes interviennent sur la même information en même temps. Sans entrer dans la complexité des recherches menées dans le domaine du travail collaboratif, ni dans les développements émergeant dans le monde XML, il est important de constater que l'ingénierie linguistique a un rôle non négligeable à jouer dans le domaine.

Notre objectif est de décrire les principales caractéristiques d'un environnement d'édition en ligne de terminologies multilingues[6], telles que manipulées à l'heure actuelle dans des contextes de traduction ou de rédaction de documents techniques. Ces terminologies correspondent aux vocabulaires utilisés dans des domaines spécialisés (documents légaux, manuels techniques, ouvrages d'enseignement) et qui de ce fait sont supposés faire preuve d'une plus grande régularité dans leurs usages. La gestion des équivalents des termes correspondants dans différentes langues fait par ailleurs l'objet d'un contrôle plus étroit que dans le cas de dictionnaires multilingues d'usage courant de par le caractère éventuellement officiel des usages qui peuvent en être fait (par exemple dans le cadre des textes légaux de l'union européenne).

8.4.2. Utilisateurs et scénarios d'interaction

La première étape d'un projet de ce type consiste à identifier les différents acteurs qui vont intervenir et les scénarios d'interaction correspondants. La figure 8.4 montre une organisation éditoriale possible pour l'édition de terminologies multilingues. On y distingue ainsi :

- Les rédacteurs, qui sont, pour les différentes langues concernées, des spécialistes du domaine considéré. Ils ne sont pour autant pas des experts en terminologie et devront être guidés dans la conception des entrées multilingues par un cadre éditorial relativement rigide. Ils travaillent sur une base temporaire et discutent des évolutions possibles de la base (création d'une nouvelle entrée, compatibilités inter linguistiques etc.) par le biais d'un forum de discussion électronique (le *journal de bord*) ;
- L'administrateur de la base. Il assure la coordination des rédacteurs et valide à intervalles réguliers les entrées qui ont été uniformément renseignées. C'est en quelque sorte lui qui est le garant de la certification des données offertes aux utilisateurs finaux ;
- Le lecteur, qui est ici un spécialiste (traducteur, rédacteur technique ou enseignant) susceptible non seulement de consulter la base de donnée terminologique validée, mais aussi de fournir un retour d'expérience lié à sa propre expertise pour suggérer la création ou la modification d'une entrée. On lui donne ainsi la possibilité d'intervenir au niveau du journal de bord.

[6] Les travaux présentés ici sont principalement issus du projet MLIS-DHYDRO, soutenu par l'union européenne (http://www.loria.fr/projets/MLIS/Dhydro).

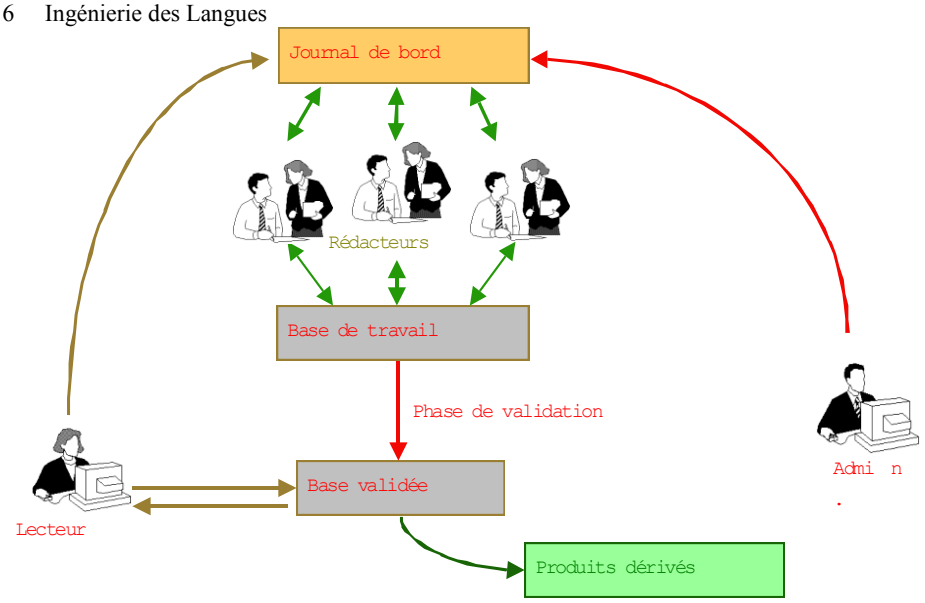


Figure 8.4 : *Organisation de la structure éditoriale*

8.4.3. Modèle de données

Les dictionnaires multilingues classiques, tout comme les ouvrages monolingues correspondants, repose sur un modèle descriptif de type sémasiologique, c'est-à-dire où l'on considère les différents sens qu'il est possible d'associer à une forme donnée. Dans le cas des lexiques spécialisés, où ce qui prime est la description adéquate des sens à l'intérieur d'un domaine de spécialité particulier, il est souvent préférable de s'appuyer sur une démarche onomasiologique où, partant d'un sens vu comme un concept du domaine, on en décrit les différentes expressions possibles, c'est-à-dire la ou les formes pouvant exprimer ce concept. Transposé au cas multilingue, ce modèle permet aisément de décrire, pour un concept donné et pour chacune des langues considérées, les différentes formes considérées alors comme de possibles équivalents de traduction. Comme on le voit sur la figure 8.5, le schéma s'articule donc autour de trois niveaux principaux :

- Le niveau du concept, où sont décrites en particulier les relations avec d'autres concepts (concept générique/spécifique, relation partie-tout) ;
- Le niveau de la langue, où l'on trouvera la définition du concept pour la langue donnée (dans le cas du dictionnaire hydrographique, les rédacteurs espagnols ont fait le choix de traduire directement les définitions anglaises alors que les rédacteurs français ont reformulé celles-ci) ;

- Le niveau du terme, où sont exprimées d'une part les caractéristiques morphosyntaxiques propres au terme et d'autres part un certain nombre d'informations spécifiques d'usage (par exemple obsolescence du terme, ou exemple d'attestation) et de lien avec les autres termes (terme préféré, abréviation etc.).

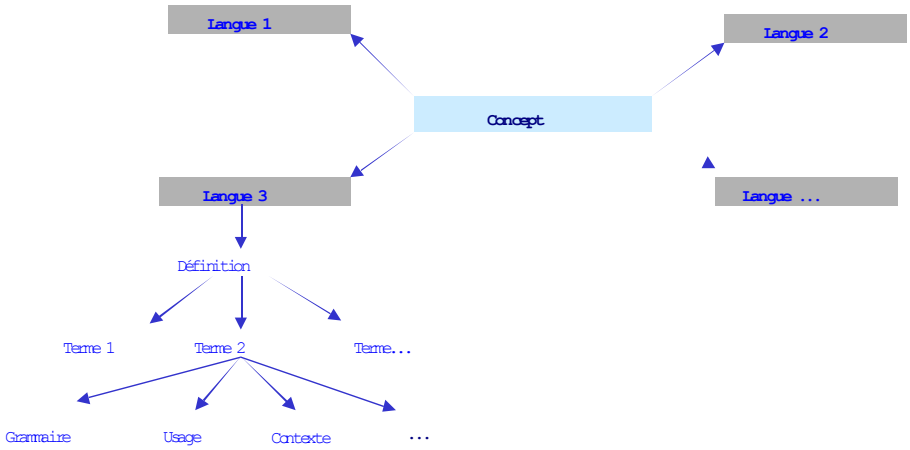


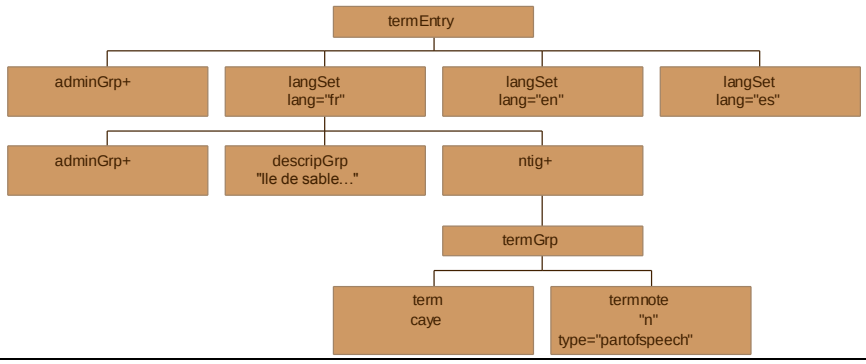
Figure 8.5 : *organisation du modèle conceptuel.*

Ce modèle est particulièrement bien adapté à l'organisation éditoriale présentée précédemment, puisqu'il est possible de considérer que chaque rédacteur est responsable, à l'intérieur d'un concept donné de l'ensemble des informations relatives à la langue dont il relève, et par ailleurs, la phase de validation s'opère effectivement au niveau global du concept. Cependant, il faut avoir conscience, d'un point de vue plus linguistique, des limitations du modèle si on voulait l'appliquer plus généralement à toute description lexicale. Il est mal adapté à gérer les phénomènes de glissement de sens et se fragmente ainsi très vite dès que la couverture linguistique envisagée devient trop large. Il faut le réserver à la gestion de données lexicales dans un ensemble de domaine spécialisés, quitte à intégrer ces bases locales, mais en parfaite connaissance de cause, à un système d'interrogation plus général dans le cadre d'études lexicographiques ou pour en récupérer les données dans un système de traitement automatique (dialogue homme-machine finalisé ou traduction automatique par exemple).

8.4.4. Implémentation

La réalisation concrète d'un système d'édition en ligne de terminologies multilingues ressemble à un certain niveau au modèle d'accès en réseau présenté à la section précédente. Les niveaux de gestion des flux de données (par exemple la connexion et

l'identification des rédacteurs ou le téléchargement des fiches terminologique par ceux-ci) peuvent être gérés de même façon à l'aide d'un format de représentation et de communication reposant sur le format XML. De ce point de vue, la représentation des données primaire sous une forme compatible avec nos formats d'échange se trouve facilitée par l'existence d'un standard international exprimé en SGML et facilement transposable à XML. La norme Martif instancie ainsi pleinement le modèle conceptuel sous la forme d'une structure générique à trois niveau tel qu'on peut l'observée figure 8.6. On y voir par ailleurs l'existence de champs particuliers (adminGrp) permettant d'adjoindre à chacun de ces niveaux des informations administratives qui permettent de gérer les phases de création et de modification des entrées, tout en contrôlant les droits d'accès de telle ou telle sous-partie du schéma à un rédacteur donné.




```
<termentry type="conceptentry" id="dhi-6">
<langset lang="fr" id="dhi-6-fr">
<descripgrp><descrip type="definition">île de sable et de corail, basse et
      plate, affleurant ou asséchant à basse mer.</descrip></descripgrp>
<ntig>
<termgrp>
<term id="dhi-6-fr-1">Caye</term>
<termnote type="termtype">main entry</termnote>
<termnote type="partofspeech">n</termnote>
<termnote type="grammaticalgender">f</termnote>
<termnote  type="timerestriction">Terme employé à l'origine pour désigner
      les îlots coralliens de la Mer des Antilles.</termnote>
</termgrp></ntig></langset></termentry>
```

Figure 8.6 : *représentation d'une entrée terminologique au Martif (schéma général et entrée simplifiée)*

Par ailleurs, on perçoit l'intérêt de XML comme format de représentation de la structure logique de document quand il s'agit de produire, du point de vue de la présentation, des documents papiers (impression à la demande) ou électroniques

(navigatio WWW ou CDROM) qui sont mis à jour au fur et à mesure de l'évolution éditoriale. Il suffit de produire des feuilles de style XSL qui fourniront autant de sorties que nécessaire à partir d'une même base validée. Ces mêmes feuilles de styles, de par leur indépendance du logiciel proprement dit, pourront d'ailleurs être maintenues ou complétées directement par les utilisateurs.

8.4.5 Bilan partiel

L'exemple d'un projet tel que Dhydro montre qu'il est possible à la fois d'appliquer des méthodologies génériques issues de recherches académiques et de se conformer aux exigences de la réalisation de prototypes opérationnels, notamment en termes de normalisation. L'une des conséquences importantes d'une telle démarche est que l'on peut enfin envisager de réutiliser un certain nombre de composants ainsi définis dans le cadre de plate-formes de recherche et échanger les données lexicales correspondantes.

8.5 Perspectives

Le présent chapitre s'est efforcé de donner un certain nombre de pistes concrètes permettant de mettre en œuvre des outils ou des plates-formes d'accès aux ressources linguistiques. On a pu ainsi observer l'absence à l'heure actuelle d'une véritable base conceptuelle pour le développement d'environnements de programmation partagés qui permettraient de développer des applications complexes sur la base de composants élémentaires normalisés. Il reste cependant que le développement de standards tels que XML, qui dépasse de loin le champ du traitement automatique des langues, est susceptible de faire progresser la discipline pour peu que de nombreuses équipes rebondissent sur l'opportunité qui est offerte. Cela peut conduire en particulier à la définition de normes de codages pour différents types de ressources, notamment lexico-syntaxiques, qui pourront être décrites à l'aide de DTD XML, et par effet d'entraînement à la conception d'outils génériques pour manipuler celles-ci.

8.6. Bibliographie

[Adda 1999] Adda G., Mariani J., Paroubek P., Rajman M. et Lecomte J. L'action GRACE d'évaluation de l'assignation des parties du discours pour le français, *Langues*, 2(2), juin 1999.

[Anderson 1991] Anderson, Anne H., Miles Bader, Bard E., Boyle E., Doherty G., Garrod S., Isard S., Kowtko J., McAllister J., Miller J., Sotillo C., Thompson H. et Weinert R. The HCRC Map Task Corpus. *Language and Speech*, 34(4), pp.351-366.

[Bel 1995] Bel N., Calzolari N. et Monachini M. (Eds.) Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets, LRE Project 62-050 Multext, Delivrable D1.6.1B.

[Bourion 1998] Bourion E. Ponctuation et accès sémantique aux banques textuelles, in Defays J.-M., Rosier L. et ikin F. (eds.) Actes du colloque *A qui la ponctuation ?* Liège, pp. 409-435.

[Crist 1994] Christ O. A modular and flexible architecture for an integrated corpus query system. *COMPLEX'94*, Budapest, 1994.

[Church 1991] Church K., Gale W., Hanks P. et Hindle D. Using Statistics in Lexical Analysis, in U. Zernik (ed.) *Lexical Acquisition : Using On-Line Resources to Built a Laxicon*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 115-163.

[Dagan 1994] Dagan I. et Church K. W. Termight: Identifying and Translating Technical Terminology. In *4th Conference on Applied Natural Language Processing*, Stuttgart, Association for Computational Linguistics, 1994.

[DOM 1998] Document Object Model (DOM) Level 1 Specification, Version 1.0, Recommandation du W3C, 1er octobre 1998. http://www.w3.org/TR/REC-DOM-Level-1/

[Elan 1999] European Language Activity Network, Deliverables of WorkPackage 3, décembre 1999. http://www.loria.fr/projets/MLIS/Elan

[Muller 1977] Muller C. *Principes et méthodes de statistique lexicale*, Champion, collection Unichamp, Paris.

[Namespaces 1999] Namespaces in XML, Recommandation du W3C, 14 janvier 1999. http://www.w3.org/TR/REC-xml-names

[Rastier 1994] Rastier F. Microsémantique, lexique et contexte. In Martin E. (Ed.) *Traitements informatisés de corpus textuels*, Didier, Paris, pp. 109-147.

[Rastier 1995] Rastier F. La sémantique des thèmes – ou le voyage sentimental. In Rastier F. (Ed.) *L'analyse thématique des données textuelles*, Didier, Paris, pp.223-249.

[SAX 1998] SAX 1.0: The Simple API for XML, 11 mai 1998. http://www.megginson.com/SAX/index.html

[Sinclair 1991] Sinclair, J. *Corpus, Concordance, Collocation*. Oxford: OUP. 1991.

[Valceschini-Deza 1999] Valcheschini-Deza N. *Accès sémantique aux bases de données textuelles*, Thèse de l'Université de Nancy 2.

[XSLT 1999] XSL Transformations (XSLT), Version 1.0, Recommandation du W3C, 16 novembre 1999. http://www.w3.org/TR/xslt

[Xpath 1999] XML Path Language (XPath), Version 1.0, Recommandation du W3C, 16 novembre 1999. http://www.w3.org/TR/xpath